



ESIP Data Citation Guidelines

Ruth Duerr
National Snow and Ice Data Center





Outline

- Purpose
- Background
- General format
- Mandatory fields
- Optional fields



Purpose of data citation

- To aid scientific **reproducibility** through direct, unambiguous reference to the precise data used in a particular study. (This is the paramount purpose and also the hardest to achieve).
- To provide fair **credit** for data creators or authors, data stewards, and other critical people in the data production and curation process.
- To ensure scientific **transparency** and reasonable **accountability** for authors and stewards.
- To aid in tracking the **impact** of data set and the associated data center through reference in scientific literature.
- To help data authors verify how their data are being **used**.
- To help future data users identify how others have **used** the data.



Background

- ESIP Preservation and Stewardship cluster has spent several years working on the identifier and citation topic
- Guideline heritage:
 - Examination of guidelines from ESIP member data centers
 - IPY citation guidelines
 - Tested during application at meetings such as GeoData 2011 and ESIP meetings
- Draft citation guidelines were approved by the General Assembly at Jan. 2012 winter meeting



General format

- Cite a data set as if it were a book!
- For example,
Cline, D., R. Armstrong, R. Davis, K. Elder, and G. Liston. 2002, Updated 2003. CLPX-Ground: ISA snow depth transects and related measurements ver. 2.0. Edited by M. Parsons and M. J. Brodzik. National Snow and Ice Data Center. Data set accessed 2008-05-14 at <http://dx.doi.org/10.5060/D4MW2F23z>



Mandatory Fields

- **Author(s)**--the people or organizations responsible for the intellectual work to develop the data set. The data creators.
- **Release Date**--when the particular version of the data set was first made available for use (and potential citation) by others.
- **Title**--the formal title of the data set
- **Version**--the precise version of the data used. Careful version tracking is critical to accurate citation.



Mandatory Fields

- **Archive and/or Distributor**--the organization distributing or caring for the data, ideally over the long term.
- **Locator/Identifier/Distribution Medium**--this could be a URL but ideally it should be a persistent service, such as a DOI, Handle or ARK, that resolves to the current location of the data in question.
- **Access Date and Time**--because data can be dynamic and changeable in ways that are not always reflected in release dates and versions, it is important to indicate when on-line data were accessed.



Author

- Data stewards should work with data providers to determine who gets credit and takes responsibility for the data set
- Small groups and organizations may be authors
 - Be as specific as possible for accountability and credit purposes

The FOO Working Group. 2001. The FOO Data Set. The FOO Data Center. <http://dx.doi.org/10.xxxx/notfoo.547983>. Accessed 1 May 2011. Subsets based on sub-collections or representation of the data set



Author (continued)

- A data set that is a collection of several smaller, independent data sets will not have an author
 - The individual data sets would have their own specific citations with author
 - The collection would likely have an editor or compiler, though:

Doe, J. (compiler) 2001. The FOO Collection. The FOO Data Center. <http://dx.doi.org/10.xxxx/notfoo.547983>. Accessed 1 May 2011.



Release date

- For a completed data set, the release date is simply the year of release.

Doe, J. and R. Roe. 2001. The FOO Data Set. The FOO Data Center. <http://dx.doi.org/10.xxxx/notfoo.547983>. Accessed 1 May 2011.

- More precise dates can be used if needed to indicate exactly when the data became available and citable



Release date (continued)

- If detailed versioning information is lacking, try and capture when updates occurred.
 - For infrequently or irregularly updated data, list the first year of released followed by "updated" with the current update information.

Doe, J. and R. Roe. 2001, updated 2005. The FOO Occasionally Updated Data Set. The FOO Data Center. <http://dx.doi.org/10.xxxx/notfoo.547983>. Accessed 1 May 2011.

- For an ongoing and regularly updated data set, list the first year of release followed by the last update.

Doe, J. and R. Roe. 2001, updated daily. The FOO Time Series Data Set. The FOO Data Center. <http://dx.doi.org/10.xxxx/notfoo.547983>. Accessed 1 May 2011.



Title

Best Practices

- Try to avoid having version or edition information in the title.
- Should not be the title of a project or a related publication.
- A data set should have an identity and title of its own.



Version

- Careful versioning and documentation of version changes are central to enabling accurate citation.
- Include version as part of the citation for any version greater than 1.
- It may be appropriate to track major and minor versions.

Doe, J. and R. Roe. 2001. The FOO Data Set. Version 2.3. The FOO Data Center. <http://dx.doi.org/10.xxxx/notfoo.547983>. Accessed 1 May 2011.



Archive and/or distributor

- The organization that maintains and manages the release or distribution of the data set.
 - Often implies responsibility for stewardship of the data set.
 - Often considered the data "publisher."
 - DataCite describes this role as:

"The entity that holds, archives, publishes, prints, distributes, releases, issues, or produces the resource. This property will be used to formulate the citation, so consider the prominence of the role."
- May be appropriate to recognize a major sponsor of the data here.

Doe, J. and R. Roe. 2001. The FOO Data Set. The FOO Funding Agency Data Center. <http://dx.doi.org/10.xxxx/notfoo.547983>. Accessed 1 May 2011.



Locator, Identifier, or Distribution Medium

- If there is one fixed distribution medium, include it - DVD or CD-ROM
- If the data is available on the internet or via multiple media a persistent locator is required
 - PURLs, ARKs, DOIs, Handles, or any other persistent equivalent will do
 - DOIs are favored by publishers
 - Thompson Reuter is working on including data sets with DOIs and possibly ARKs in the Web of Science



Locator, Identifier (continued)

- Best practices
 - Locators should point to a landing page for the data set
 - Do not include the name of an organization in the locator
 - Use the http form of the locator for human usability

Doe, J. and R. Roe. 2001. The FOO Data Set. Version 2.3. The FOO Data Center. <http://dx.doi.org/10.xxxx/notfoo.547983>. Accessed 1 May 2011.



Optional Fields

- **Subset Used** - The logical equivalent of citing a passage in a book
- **Editor, Compiler, or other important role** - other roles may need to be recognized especially if there is no author
- **Archive or Distributor Place** - the city, state (if necessary), and country of the archive or distributor
- **Distributor, Associate Archive, or other Institutional Role**
- **Data within a larger work**



Subset used

- Data stewards should suggest reasonable ways to specify a subset of a data set
- Specifying a temporal and/or spatial subset
Doe, J. and R. Roe. 2001, updated daily. The FOO Gridded Time Series Data Set. Version 3.2. Oct. 2007- Sep. 2008, 84°N, 75°W; 44°N, 10°W. The FOO Data Center. <http://dx.doi.org/10.xxxx/notfoo.547983>. Accessed 1 May 2011.
- Subsets based on sub-collections or representations of the data set
Doe, J. and R. Roe. 2001. The FOO Data Set. Version 2.0 shapefiles. The FOO Data Center. <http://dx.doi.org/10.xxxx/notfoo.547983>. Accessed 1 May 2011.
- In the future it is conceivable that systems may be able to record the exact data that corresponds to a particular data access query



Editor, Compiler, or other important role

- **Editor** - a person or team who is responsible for creating a value-added and possibly quality-controlled product from the data.
- **Compiler** - a person who is responsible for compiling a product from the data albeit with minimal scientific or technical input.

Doe, J. 2001. The FOO Data Set. Version 2.0 R. Roe (ed.)
The FOO Data Center. <http://dx.doi.org/10.xxxx/notfoo.547983>. Accessed 1 May 2011.



Editor, Compiler, or other important role (cont.)

- Editors and compilers may often be responsible for a larger work that includes multiple data sets from different authors

Doe, J. (ed.). 2001. The FOO Data Set. Version 2.0 The FOO Data Center. <http://dx.doi.org/10.xxxx/notfoo.547983>. Accessed 1 May 2011.



Editor, Compiler, or other important role (cont.)

- Occasionally there may be both a compiler and editor as well as other roles
 - Only acknowledge at most two roles in the citation
 - Record other roles in metadata or documentation

Doe, J. (ed.). 2001. The FOO Data Set. Version 2.0 R. Roe (compiler) The FOO Data Center. <http://dx.doi.org/10.xxxx/notfoo.547983>. Accessed 1 May 2011.

Distributor, associate archive, or other institutional role



- There may be multiple organizational roles that need to be recognized
 - This can be a way to recognize a funding source

Doe, J. 2001. The FOO Data Set. Version 2.0 The FOO Data Center. Distributed by the FEU Distribution Center. <http://dx.doi.org/10.xxxx/notfoo.547983>. Accessed 1 May 2011.

Doe, J. 2001. The FOO Data Set. Version 2.0 The FOO Data Center in association with the FUU Data Center. <http://dx.doi.org/10.xxxx/notfoo.547983>. Accessed 1 May 2011.



Data within a larger work

- Cite data sets within a larger work similarly to how a chapter in a book is cited

Bockheim, J. 2003. "University of Wisconsin Antarctic Soils Database". In International Permafrost Association Standing Committee on Data Information and Communication (comp.). 2003. Circumpolar Active-Layer Permafrost System, Version 2.0. Edited by M. Parsons and T. Zhang. Boulder, CO: National Snow and Ice Data Center/World Data Center for Glaciology. CD-ROM.